Note on Existence and Non-Existence of Large Subsets of Binary Vectors with Similar Distances

Gregory Gutin and Mark Jones
Royal Holloway, University of London
Egham, Surrey, TW20 0EX, UK

February 29, 2012

Abstract

We consider vectors from $\{0,1\}^n$. The weight of such a vector v is the sum of the coordinates of v. The distance ratio of a set L of vectors is $\mathrm{dr}(L) := \max\{\rho(x,y):\ x,y \in L\}/\min\{\rho(x,y):\ x,y \in L,\ x \neq y\}$, where $\rho(x,y)$ is the Hamming distance between x and y. We prove that (a) there are no positive constants α and C such that every set K of vectors with weight p contains a subset K' with $|K'| \geq |K|^\alpha$ and $\mathrm{dr}(K') \leq C$, even when $|K| \geq 2^p$, (b) for a set K of vectors with weight p, and a constant $C > 2$, there exists $K' \subseteq K$ such that $\mathrm{dr}(K) \leq C$ and $K' \geq |K|^\alpha$, where $\alpha = 1/\lceil \log(p/2)/\log(C/2)\rceil$.

1 Introduction

We will consider n-dimensional binary vectors (i.e., vectors from $\{0,1\}^n$) and call them *n-vectors*. The *(Hamming) weight* $|v|$ of an n-vector v is the sum of the coordinates of v. The *(Hamming) distance* $\rho(u,v)$ between n-vectors u, v is the number of coordinates where u and v differ. The *distance ratio* of a set L of n-vectors is

$$\mathrm{dr}(L) := \frac{\max\{\rho(x,y):\ x,y \in L\}}{\min\{\rho(x,y):\ x,y \in L,\ x \neq y\}}.$$

Let $p \leq n$ be positive integers. Abramovich and Grinshtein [1] asked whether the following claim holds true:

Claim 1. *There exist positive constants α and C such that every set K of at least 2^p n-vectors with Hamming weight p contains a subset K' with $|K'| \geq |K|^\alpha$ and* $\mathrm{dr}(K') \leq C$.

If the claim is true, it can be used in statistics for establishing the lower bounds for the minimax risk of estimation in various sparse settings [1, 3]. If the claim is not true, a counterexample can be used to impose some conditions on K such that the claim becomes true and, thus, is still useful for establishing minimax lower bounds over narrower classes of settings. Also, a weaker bound on $|K'|$ can be used to obtain weaker lower bounds for the risk of estimation.

The following example shows that for some sets K the claim is true. Let $p < n/2$ and let Ω denote the set of all n-vectors of weight p. By Lemma A.3

in [3] (which is a generalization of the Varshamov-Gilbert lemma attributed to Reynaud-Bouret [2]), there exists a subset Ω' of Ω such that $\rho(x, y) \geq (p+1)/4$ for all distinct $x, y \in \Omega'$ and $|\Omega'| \geq (1 + en/p)^{\beta p}$ for some $\beta \geq 9 \cdot 10^{-4}$. It follows that $\mathrm{dr}(\Omega') < 8$ (since $\rho(u, v) \leq 2p$ for all $u, v \in \Omega$). Moreover, since $|\Omega| = \binom{n}{p} < (en/p)^p$ and $|\Omega'| > (en/p)^{\beta p}$, we have $|\Omega'| > |\Omega|^{\beta}$. For sufficiently large n/p, $|\Omega'| \geq 2^p$.

Unfortunately, in general, the claim is not true and we give a counterexample to the claim in Section 2. In Section 3, we show that a weaker claim holds: there exists $K' \subseteq K$ such that $\mathrm{dr}(K) \leq C$ and $K' \geq |K|^{\alpha}$, where $\alpha = 1/\lceil \log(p/2)/\log(C/2) \rceil$ $(C > 2)$.

Henceforth $[s] := \{1, \ldots, s\}$ for a positive integer s.

2 Counterexample

Let us fix constants $C \geq 1$ and $0 < \alpha \leq 1$. We will show that there is no set K of n-vectors satisfying Claim 1 for these C and α. In this section, we will use fixed positive integers t, a, p, q and n satisfying the following:

1. $1/t < \alpha$, $a > C$;
2. p is a multiple of a^t;
3. $q^t \geq 2^p$;
4. $n \geq p + p(q-1)\sum_{j=1}^{j=t}(q/a)^{t-j}$.

We say a set L of n-vectors is a $\mathcal{C}_0$*-set* if L consists of a single vector. For $i \in [t]$, a set L of vectors is a $\mathcal{C}_i$*-set* if it satisfies the following:

1. $|L| = q^i$;
2. $\max\{\rho(x, y) : \ x, y \in L\} = 2p/a^{t-i}$;
3. L can be partitioned into q sets $L_1, \ldots, L_q$ such that for each r, L_r is a $\mathcal{C}_{i-1}$-set, and for all $x \in L_r, y \in L_s$ with $r \neq s$, $\rho(x, y) = 2p/a^{t-i}$.

Lemma 1. *For each $i \in [t]$, there is a set K of n-vectors such that K is a $\mathcal{C}_i$-set.*

Proof. For a set L of n-vectors to be a $\mathcal{C}_i$-set, we need that

$$\max\{\rho(x, y) : \ x, y \in L\} = 2p/a^{t-i}.$$

So for every pair $x, y \in L$ of distinct n-vectors, there must be a set $X \subseteq [n]$ with $|X| \geq p - p/a^{t-i}$, such that $x_i = y_i = 1$ for all $i \in X$. In fact, in our construction below we will assure that in a $\mathcal{C}_i$-set, there exists $X \subseteq [n]$ with $|X| \geq p - p/a^{t-i}$ such that $x_i = 1$ for all $x \in L$.

For some $S \subseteq T \subseteq [n]$, we say a set L of n-vectors is a $\mathcal{C}_i$-set *between* (S, T) if L is a $\mathcal{C}_i$-set, and for all $x \in L$, $x_r = 1$ if $r \in S$ and $x_r = 0$ if $r \notin T$. We give a recursive method to construct a $\mathcal{C}_i$-set between (S, T) when $|S| = p - p/a^{t-i}$ and $|T|$ is large enough (we calculate the required size of T later). We can then construct the required set K by constructing a $\mathcal{C}_t$-set between $(\emptyset, [n])$.

Given S, T, construct a $\mathcal{C}_i$-set L between (S, T) as follows. If $i = 0$, return a single n-vector x of Hamming weight p, such that $x_r = 1$ for all $r \in S$ and $x_r = 0$ for all $r \notin T$.

If $i \geq 1$, partition $T \backslash S$ into q sets $T_1, \ldots, T_q$, such that $-1 \leq |T_r| - |T_s| \leq 1$ for all r, s. For each $1 \leq r \leq q$, let S_r be a subset of T_r of size $p/a^{t-i} - p/a^{t-(i-1)}$. Then for each r construct a $\mathcal{C}_{i-1}$-set L_r between $(S \cup S_r, S \cup T_r)$, and let L be the union of these sets. (Note that $|S \cup S_r| = p - p/a^{t-(i-1)}$, as required for the recursion.)

Observe that since $|S| = p - p/a^{t-i}$, $\max\{\rho(x, y) : \ x, y \in L\} \leq 2p/a^{t-i}$. Furthermore, since $T_1, \ldots, T_q$ are disjoint, for $x \in L_r, y \in L_s$ with $r \neq s$, $\rho(x, y) = 2p/a^{t-i}$. Finally note that $|L| = \sum_{r=1}^{r=q} |L_r| = qq^{i-1} = q^i$. Therefore L satisfies all the conditions of a $\mathcal{C}_i$-set between (S, T).

We now calculate a bound f_i such that we can construct a $\mathcal{C}_i$-set between (S, T) when $|S| = p - p/a^{t-i}$ as long as $|T| \geq f_i$.

Clearly $f_0 = p$. For $i > 0$, in the construction above we require that $|S \cup T_r| \geq f_{i-1}$ for each $1 \leq r \leq q$. Therefore we require

$$f_i = |S| + q(f_{i-1} - |S|) = qf_{i-1} - (q-1)(p - p/a^{t-i}).$$

Observe that this is satisfied by setting $f_i = p + p(q-1) \sum_{j=1}^{j=i} (q^{i-j}/a^{t-j})$.

So to construct a $\mathcal{C}_i$-set between $(\emptyset, [n])$, it suffices that $n \geq p + p(q-1) \sum_{j=1}^{j=i} (q/a)^{t-j}$, which holds by Part 4 of the conditions on t, a, p, q and n given in the beginning of this section. □

Theorem 1. *There is a set K of n-vectors for which Claim 1 does not hold.*

Proof. We will construct a set K such that for any subset of K with more than $q = |K|^{1/t}$ vectors, the distance ratio is at least a. This implies that for any subset with at least $|K|^\alpha$ vectors the distance ratio is greater than C, as required.

By Lemma 1, we may assume that we have a $\mathcal{C}_i$-set K. Thus, K can be partitioned into q sets $K_1, \ldots, K_q$ such that for each r, K_r is a $\mathcal{C}_{i-1}$-set, and for all $x \in K_r, y \in K_s$ with $r \neq s$, $\rho(x, y) = 2p/a^{t-i}$.

Note that any subset $K' \subseteq K$ of more than q vectors will contain at least two vectors from K_r for some r and so $\min\{\rho(x, y) : \ x, y \in K', x \neq y\} \leq 2p/a^{t-i+1}$; furthermore if K' contains vectors from K_r and K_s for $r \neq s$ then $\max\{\rho(x, y) : x, y \in K'\} \geq 2p/a^{t-i}$.

Therefore, for any $K' \subseteq K$ with $|K'| > q$, either $\mathrm{dr}(K') \geq a$, or $K' \subseteq K_i$ for some $\mathcal{C}_{i-1}$-set K_i. Furthermore there is no $K' \subseteq K$ with $|K'| > q$ if K is a $\mathcal{C}_1$-set. So by induction on $i \geq 1$, every $K' \subseteq K$ with $|K'| > q$ has $\mathrm{dr}(K') \geq a$. By letting $i = t$, we complete the proof of the theorem. □

3 Positive Result

Given a set K of n-vectors, we are interested in finding a subset $K' \subseteq K$ as large as possible such that $\mathrm{dr}(K') \leq C$, for some constant C. The following is such a result.

Theorem 2. *Let K be a set of n-vectors with Hamming weight exactly p, and let $C > 2$ be a constant. Then there exists $K' \subseteq K$ such that $\mathrm{dr}(K) \leq C$ and $K' \geq |K|^\alpha$, where $\alpha = 1/\lceil \log(p/2)/\log(C/2) \rceil$.*

Proof. Let $t = \lceil \log(p/2)/\log(C/2) \rceil = 1/\alpha$.

Let $K_1 = K$. For each $1 \leq i < t$, let K_{i+1} be a maximal subset of K_i such that $\min\{\rho(x,y)|x,y \in K_{i+1},\ x \neq y\} \geq C^i/2^{i-1}$. For each vector $z \in K_i$, let $N_i(z)$ be the set of vectors $x \in K_i$ for which $\rho(x,z) \leq C^i/2^{i-1}$.

Observe that $\max\{\rho(x,y)|x,y \in N_i(z)\} \leq C^i/2^{i-2}$. Therefore since $\min\{\rho(x,y)|x,y \in K_i\} \geq C^{i-1}/2^{i-2}$, we have $\mathrm{dr}(N_i(z)) \leq C$. Note furthermore that by the maximality of K_{i+1}, every vector in K_i is in $N_i(x)$ for some $x \in K_{i+1}$. Therefore, for $1 \leq i < t$, we either have that $|N_i(x)| \geq |K|^\alpha$ for some $x \in K_{i+1}$, in which case we are done, or $|K|^\alpha |K_{i+1}| \geq |K_i|$. By induction, we have that $|K_i| \geq |K|/|K|^{\alpha(i-1)}$ for $1 \leq i \leq t$ (or else we can find a set $N_i(x)$ satisfying the theorem). In particular, we have that $|K_t| \geq |K|/|K|^{\alpha(t-1)} = |K|/|K|^{1-\alpha} = |K|^\alpha$.

Now observe that $\max\{\rho(x,y)|x,y \in K_t\} \leq 2p$. Furthermore,

$$\min\{\rho(x,y)|x,y \in K_t,\ x \neq y\} \geq C^{t-1}/2^{t-2} = (4/C)(C/2)^t \geq (4/C)(p/2) = 2p/C.$$

Therefore $\mathrm{dr}(K_t) \leq C$. This completes the proof. □

References

[1] Felix Abramovich and Vadim Grinshtein, Private communication, Jan. 2012.

[2] Patricia Reynaud-Bouret, Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. Probab. Theory Related Fields 126: 103–153, 2003.

[3] Philippe Rigollet and Alexandre Tsybakov, Exponential Screening and optimal rates of sparse estimation. Ann. Statist. 39(2): 731–771, 2011. (See also http://arxiv.org/abs/ArXiv:1003.2654v3.)